The Stata Journal

Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142; FAX 979-845-3144
jnewton@stata-journal.com

Department of Geography
University of Durham
South Road
Durham City DH1 3LE UK
n.j.cox@stata-journal.com

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2005)
5, Number 1, pp. 130–133

Review of Generalized Latent Variable Modeling by Skrondal and Rabe-Hesketh

Roger Newson
King's College London, UK
roger.newson@kcl.ac.uk

Abstract. The new book by Skrondal and Rabe-Hesketh (2004) is reviewed.

Keywords: gn0025, GLLAMM, generalized linear latent and mixed models, latent variables

1 Introduction

This is a very impressive book. The authors introduce a class of statistical models, known as generalized latent variable models (GLVMs), which contains, as a special case, the generalized linear models (GLMs) introduced by McCullagh and Nelder (1989). GLMs have the feature that the conditional mean of the outcome Y-variable, given values of a list of X-variables, can be transformed by a link function to give a linear predictor equal to a linear combination of these X-variables. The estimated parameters are typically interpreted as proportions, odds, probits, or arithmetic, geometric, harmonic, or algebraic means and their differences or ratios. This new book extends GLMs by allowing the linear predictor to contain not only terms containing X-variables whose values we actually know, but also terms containing "latent variables", or hidden variables, whose values we can only imagine. The estimated parameters are therefore typically interpreted as *conditional* proportions, odds, probits, or arithmetic, geometric, harmonic, or algebraic means (and their differences or ratios), conditioning on the values both of the observed X-variables and of the unobserved latent variables. The latent variables are usually assumed to be sampled from a distribution with a zero mean, and to have a value, in each observation, specific to a cluster or subcluster to which that observation belongs. Typically, therefore, the latent variables are cluster or subcluster effects, representing a difference between a particular cluster or subcluster and the "average" cluster or subcluster. For instance, if the Y-variable is an exam mark produced by a particular student in a particular class in a particular school on a particular day, then the observations might correspond to exam scripts, the clusters might be schools, and the subclusters might be classes within schools and students within classes. Most (but not all) of the models discussed can be fitted to data using the `gllamm` package in Stata (see Rabe-Hesketh, Skrondal, and Pickles [2004]), downloadable from SSC. The authors argue that many statistical methods, such as common factor analysis, latent class models, frailty models, random-effects models, and multilevel modeling, can be seen as special cases of GLVMs and implemented using `gllamm`. However, this book is primarily about GLVMs (the generic family of methods), and not about `gllamm` (the means for implementing these methods specifically in Stata), which will be covered more fully in a forthcoming book from Stata Press (Rabe-Hesketh, Pickles, and Skrondal 2005).

GLVMs are subtly different from generalized estimating equation (GEE) models, which are an alternative extension of GLMs to clustered data (see Hardin and Hilbe [2003]). The parameters of a GLVM corresponding to the X-variables describe the *conditional* distribution of the Y-variable, given the values both of the observed X-variables and of the unobserved latent variables. The corresponding parameters of a GEE model describe the distribution of the Y-variable, conditional on the X-values but marginal to the latent variables (assuming that these exist). If the parameters are risk differences, arithmetic mean differences, arithmetic mean ratios, geometric mean ratios, or risk ratios, the GEE parameters are the same as the corresponding GLVM parameters. However, this identity does not hold for all parameters. For instance, if the parameters are odds ratios, then the GEE parameters will tend systematically to be closer to one than the corresponding GLVM parameters. Therefore, if the observations belong to students, the clusters are schools, the Y-variable indicates whether an exam is passed, the X-variable of interest is a student nonsmoking status indicator, and a logit link function is used; then the GEE smoking-related odds-ratio estimates the factor whereby the odds of passing an exam would change in the population if the whole world changed from smokers to nonsmokers. The GLVM odds ratio on the other hand would estimate the factor by which the exam-passing rate *in each school* would change if all the students in that school were to change from smokers to nonsmokers. The latter odds would be further from one than the former and might arguably be more useful for the head of each school to know, particularly if each school head had some idea of his/her own school's baseline odds, assuming that all the students smoked. Also the GLVM parameters are more comparable to the corresponding parameters of the fixed-effects model that we would probably have used if we had data only on a small number of schools.

Latent variables are controversial because it is not often easy to measure them directly, and this might lead many people to wonder in what sense they can be said to exist. A GLVM enthusiast might argue that genotypes were once latent variables (in the days of Gregor Mendel) and that they were useful for some decades before we developed the technology to measure them directly, and that today's latent variables might similarly correspond to something that our descendants will be able to measure directly. More realistically, as the authors argue in section 1.7, latent variable models can be used to generate a very wide range of within-cluster dependence structures for the outcome. For instance, if there are subclusters nested within clusters, then outcomes in the same subcluster will be more correlated than outcomes in different subclusters within the same cluster, and hierarchial structures of that nature do not seem to be available using Stata's GEE programs. The Gauss–Markov theorem seems to imply that the better you can model the within-cluster dependence, the less wide the confidence limits will be for the same coverage probability. Latent variable models can therefore still be useful, even if we do not seriously believe in the latent variables themselves. The `gllamm` package offers the option of Huber variances for users who are skeptical of the possibility of guessing the correlation structure right the first time.

2 Summary

The book is divided into two parts, of which the first presents the methods and the second gives us a tour of a wide range of possible applications.

In the *Methodology* part, chapter 1 gives a survey of the uses to which latent variable models have been put (rightly or wrongly) in diverse areas of science. Chapter 2 gives a brief survey of modeling methods used for various outcomes, with particular reference to GLMs. Chapter 3 surveys a range of existing latent variable methods. Chapter 4 presents the GLVM family, which includes most of the latent variable models in the previous chapters as special cases. Chapter 5 presents analytical methods for establishing, for a particular GLVM, the feasibility of identifying those parameters which govern the distribution of the latent variables. As the authors state, these parameters are identifiable only through their effects on the conditional covariance of the Y-variable in observations belonging to the same cluster, which can be measured whether or not we believe in the latent variables themselves. Chapter 6 surveys estimation methods for the parameters, including maximum likelihood, quasilikelihood, and Bayesian methods, with particular reference to numerical integration or summation over the ranges of the hypothesized latent variables, which is important because it accounts for the high level of computing resources required by GLVMs. Chapter 7 presents methods whereby, in some circumstances, the controversial latent variables might actually be measured. These include maximum likelihood methods (which are most useful if the clusters are large and which estimate the latent variables as we might estimate fixed effects) and empirical Bayesian methods (which are most useful if we have a large number of small clusters and which explicitly pool information from all the clusters to calculate credible intervals for the cluster-specific latent variables for each individual cluster). Finally, chapter 8 gives a survey of inference, diagnostics, and goodness-of-fit tests, including a discussion of Huber variances.

In the *Applications* part, the authors give a range of examples, based on datasets that users can download. These examples are grouped by the type of the outcome variable, which in these examples may be dichotomous, ordinal, counts, survival data, polytomous discrete-choice responses, or a mixture of different response types.

3 Limitations

This is an excellent book and has fewer limitations than anybody is entitled to expect, given that the authors and their coworkers have developed a very comprehensive grand unified method mostly in the last few years. However, from the *Applications* part, it is clear that the main problem with GLVMs is going to be explaining the parameters, and sometimes the controversial latent variables, to nontechnical people. Some latent variables are more controversial than others. Most people in the medical sector have no problem believing in the counterfactual outcomes that a patient *would* have experienced, had the patient been allocated to treatments A or B, and we can estimate at least the *mean* difference between these two counterfactuals using a randomized controlled trial (see Rosenbaum [2003]). At the other end of the spectrum, the polytomous discrete-

choice models with latent utilities demonstrated in chapter 13 do not seem to correspond to any mechanism whereby I myself choose a coffee maker or an elected representative, and it does not surprise me that these models lead to the counterintuitive "paradoxes" that the authors mention. However, some of the estimated model parameters may still be useful to know. Although this book is intended mainly for readers who can understand natural logarithms and standard errors, such people usually have to explain their output to less-technical people, who (in my experience) usually understand odds ratios with confidence limits better than they understand log odds ratios with standard errors, as presented in this book. On a minor point of accuracy, it is not strictly true (as the authors suggest in section 9.5.2) that a fixed-effect meta-analysis requires the assumption of an equal treatment effect in all studies in order to be valid. If a common treatment effect is estimated using unclustered full Huber variances, then the confidence interval is a valid interval estimate of the "weighted average" treatment that *would* have been observed, *if only* we could scale up the size of each study by the same large factor, and this may be useful to know, even if the effect varies from study to study. However, I have no hesitation in recommending readers to buy this book, and look forward to seeing the forthcoming book on `gllamm` (Rabe-Hesketh, Pickles, and Skrondal 2005).

4 References

Hardin, J. W. and J. M. Hilbe. 2003. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Rabe-Hesketh, S., A. Pickles, and A. Skrondal. 2005 [forthcoming]. *Multilevel and Structural Equation Modeling for Continuous, Categorical, and Event Data*. College Station, TX: Stata Press.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2004. GLLAMM Manual. Working paper 160, University of California Berkeley, Division of Biostatistics. *http://www.bepress.com/ucbbiostat/paper160/*

Rosenbaum, P. R. 2003. *Observational Studies*. 2nd ed. New York: Springer.

Skrondal, A. and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

About the Author

Roger Newson is a Lecturer in Medical Statistics at King's College London, UK, working principally in asthma research. He wrote the Stata 5 package `rglm` for calculating semi-Huber and full Huber variances for generalized linear models.